DOCUMENT RESUME

ED 463 747 IR 021 169

AUTHOR Harvey, Anne L.; Way, Walter D.
TITLE A Comparison of Web-Based Standard Setting and Monitored Standard Setting.
PUB DATE 1999-04-00
NOTE 27p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Montreal, Quebec, Canada, April 20-22, 1999).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Computer Assisted Testing; Computer Mediated Communication; Computer Oriented Programs; Data Collection; *Evaluation Methods; Internet; Standards; *Tests; *World Wide Web
IDENTIFIERS *Standard Scores; Standard Setting

ABSTRACT

Standard setting, when carefully done, can be an expensive and time-consuming process. The modified Angoff method and the benchmark method, as utilized in this study, employ representative panels of judges to provide recommended passing scores to standard setting decision-makers. It has been considered preferable to have the judges meet in a central location to complete the standard setting and take part in the discussion. The use of a central location, however, adds travel costs and time. The World Wide Web allows for training, discussion, and data collection without requiring judges and the measurement of professionals running the study (facilitators) to travel to a central location. A Web-based standard setting system was developed to offset the costs of travel for a typical, monitored standard setting, and to improve the standardization of the training materials. A study using the Web-based system was compared to a similar study using a monitored session. The results of this study are generally favorable toward using a Web-based standard setting system, and suggest that recommended passing scores from an Internet study will be similar to those from a monitored study. A future version of the system should improve on the communication between the judges, and between the judges and facilitator. However, the teachers in this study felt similarly about the overall experience, regardless of whether it was an Internet or face-to-face study. (AEF)

A Comparison of Web-Based Standard Setting and Monitored Standard Setting

Anne L. Harvey
Walter D. Way

Educational Testing Service

Paper presented at the annual meeting of the
National Council on Measurement in Education, Montreal, April 1999

A Comparison of Web-Based Standard Setting and Monitored Standard Setting

Abstract

A web-based standard setting system was developed to offset the costs of travel for a typical, monitored standard setting, and to improve the standardization of the training materials. A study using the web-based system was compared to a similar study using a monitored session. The results of this study suggest that recommended passing scores from an Internet study will be similar to those from a monitored study. A future version of the system should improve on the communication between the judges, and between the judges and facilitator. However, the teachers in this study felt similarly about the overall experience, regardless of whether it was an Internet or a face-to-face study.

A Comparison of Web Based Standard Setting and Monitored Standard Setting

Introduction

Occasions frequently arise when a test is used to ascertain that a minimum level of knowledge is present for certification or for placement into a course or program. On those occasions, it is necessary to set a minimum score as passing. When test taker data is not available, the Angoff[1] method (1971), or one of the modified Angoff methods, of standard setting is usually the method of choice when items are scored right/wrong (Berk, 1986; Crocker & Zeiky, 1995). The benchmark method or the pass/fail method (described in Faggen, 1994) are often used for items with multiple scoring levels, such as essays.

Standard setting, when carefully done, can be an expensive and time-consuming process. The modified Angoff method and the benchmark method, as utilized in this study, employ representative panels of judges to provide recommended passing scores to standard setting decision-makers. For example, a panel of teachers representative of those teaching in a particular state will provide a recommended passing score on teacher certification tests to that state's Board of Education.

One of the hallmarks of careful standard setting is the use of discussion to assist the judges in conceptualizing a common test taker of interest (usually the minimally knowledgeable test taker) and determining the factors that will affect item difficulty for this group. To that end, it has been considered preferable to have the judges meet in a central location to complete the standard setting and take part in the discussion. The use of a central location, however, adds travel costs and can often involve an overnight stay, limiting the number of judges who may participate.

The World Wide Web computer Internet allows for training, discussion, and data collection without requiring judges and the measurement professionals running the study (facilitators) to travel to a central location. While training materials and data collection forms can be sent through the mail to remote locations, the Internet provides for discussion of the minimally knowledgeable test taker and ratings of specific items. In addition, data collection forms can be set up to write to a database, eliminating the need for a separate effort to record responses for analysis.

Factors to Consider in Web-Based Standard Setting

There are several factors to consider before using the Internet for standard setting, particularly for high stakes testing programs.

---

[1] Angoff ascribes the original idea for his popular method to L. R Tucker, c. 1952.

Cost. An obvious reason for using the Internet is to reduce travel costs for judges and/or facilitators. For national, or even state-wide, panels this cost can be considerable. In addition, for larger studies, there may be some savings in personnel needed to handle materials and data entry.

Offsetting some of these savings is the initial development costs of the system, ongoing maintenance costs, and the cost of setting up items and training materials for each new study. One point to note, however, is that the only additional cost for an additional judge is whatever the judge is paid. Other than the stipend paid, a study with 10 judges will cost no more than a study with 20 judges.

Standardization of the training materials. Standard setting panels are trained to recognize the aspects of items that make them more or less difficult, formulate a definition of the test taker of interest, and appropriately apply the method used. The quality of the training could influence both the quality of the judges' response and the judges' perception of the fairness of the final result. Displaying training materials in a written format allows for careful review and editing to ensure accurate, complete, and understandable material. The quality of a face-to-face training session may vary considerably with the experience and natural ability of the panel facilitator. On the other hand, a facilitator who is present on-site may be able to adapt more quickly to unusual requests or situations.

Assuring complete results. If the standard setting study is even moderately complex, the paperwork involved can be tremendous. Judges may fill out forms such as non-disclosure agreements, background summaries (for example, relevant degrees and experience), general and item-specific comment sheets, training evaluations, content review, and fairness review, in addition to the forms used for the particular standard setting methods employed. The computer can be utilized to ensure that the paperwork is complete and, to some extent, accurate. For example, the computer can check that all the questions on a biographical questionnaire are filled in. Data can be immediately available for analysis without scanning or hand entry of responses. At the same time, it can be easier to review a written set of forms for reasonable responses than it is to check the same set of forms online.

Presentation of just-in-time responses. Face-to-face training can provide more immediate answers to judges' questions than can usually be achieved in a web-based system. Unless the facilitator is continuously available on the web site, or the web sessions are scheduled in advance, answers will wait until the facilitator checks for questions.

The advantage of the computer is that just-in-time responses can be utilized to respond to a particular judge's work. For example, a judge who has chosen an Angoff rating of 10% can be immediately reminded, via a pop-up message, that this response is below the chance level for a particular item.

Group process and discussion. Judges should have an opportunity to engage in appropriate dialogue when formulating their concept of the minimally knowledgeable test taker and when they are completing preliminary estimates of item difficulty. Several researchers have written that discussion has an effect on both intra- and inter-rater consistency (Cizek & Fitzgerald, 1996; Fitzpatrick, 1989; Hambleton & Plake, 1995; Plake & Impara, 1996). The concern is that discussion will be curtailed by the logistics of communication over the web, thereby affecting the consistency of the judges' ratings.

On the other hand, discussion on the web has a couple of advantages. The first advantage is that a written form of communication can promote a more thoughtful exchange than a verbal message. Without the time pressure of "if I don't say it now the moment will be gone," judges can think more carefully about what they will say. Also, judges may feel more comfortable expressing a dissenting opinion when they are not in a room together. This may promote a freer exchange of opinions.

Recruiting judges and selection bias. One concern when using the Internet for standard setting is the ability to match qualified judges with the right computer equipment. While the Internet is accessible at most libraries, schools, and colleges, it is not universally easy to get online at particular times, or over a series of days. This requirement may introduce a selection bias in the panel of judges. The issue of accessibility is also present, however, when requiring judges to take time from work and family to attend a standard setting session, particularly when judges must travel some distance, and perhaps stay overnight. This requirement introduces its own bias.

An additional factor to consider is the ability for judges to interact with others in their profession. Many of the ETS essay readers and committee members freely admit that they participate largely because of the opportunity to exchange information with others in their field. While the Internet allows for some of this interaction, it may not be as immediate and satisfying as an actual face-to-face meeting.

Work Environment. In a face-to-face meeting, the facilitator has a pretty good idea of the environmental factors influencing the standard setting. Was the room too cold? Too hot? Were there distracting noises or frequent interruptions? When the Internet is used, the facilitator cannot be sure of a particular judge's working environment. Is the television on in the background, for instance? At the same time, familiar surroundings, particularly a computer study area, may allow for more careful study of the training materials. Conditions are certain to vary.

Security of test materials. Tests for which there is a standard setting study are likely to be secure tests; the items and answers are to be kept a secret from test

takers. This presents security issues for standard setting facilitators. When conducting a standard setting in a monitored environment the facilitator has some control over the distribution of materials. All materials can be checked out to a judge when they enter the site and carefully accounted for before the judge leaves the site.

When the Internet is used, a non-disclosure statement and other contracts are all that protect secure testing materials. There is nothing to prevent a judge from printing all of the test materials and distributing them at will. On the other hand, the materials are certainly no less secure than when they are placed in the hands of test administrators at testing time. The judge must be trusted to follow directions on the securing of the work site and the closing of the web site to prevent others from using the materials.

While distributing materials over the Internet, security of the transmission is certainly an issue. For sensitive testing materials using the same encryption techniques used by banks is a must. These encryption techniques protect the materials in transit and the materials are at least as secure as they would be if they were sent through the mail, a regular practice for large testing programs.

Purpose of the Study

This study will determine whether a web-based standard setting study will result in recommended passing scores that are comparable to passing scores resulting from a traditional, monitored standard setting study. In addition, the paper will explore the judges' perceptions of some of the less tangible issues, such as the group process, the working environment, and the quality of the training.

The Web-Based Standard Setting System

The web-based standard setting system used in this study is the prototype for a final system to be built later this year. The system has two modes, one for the judge and one for the facilitator. The facilitator has the following capabilities:

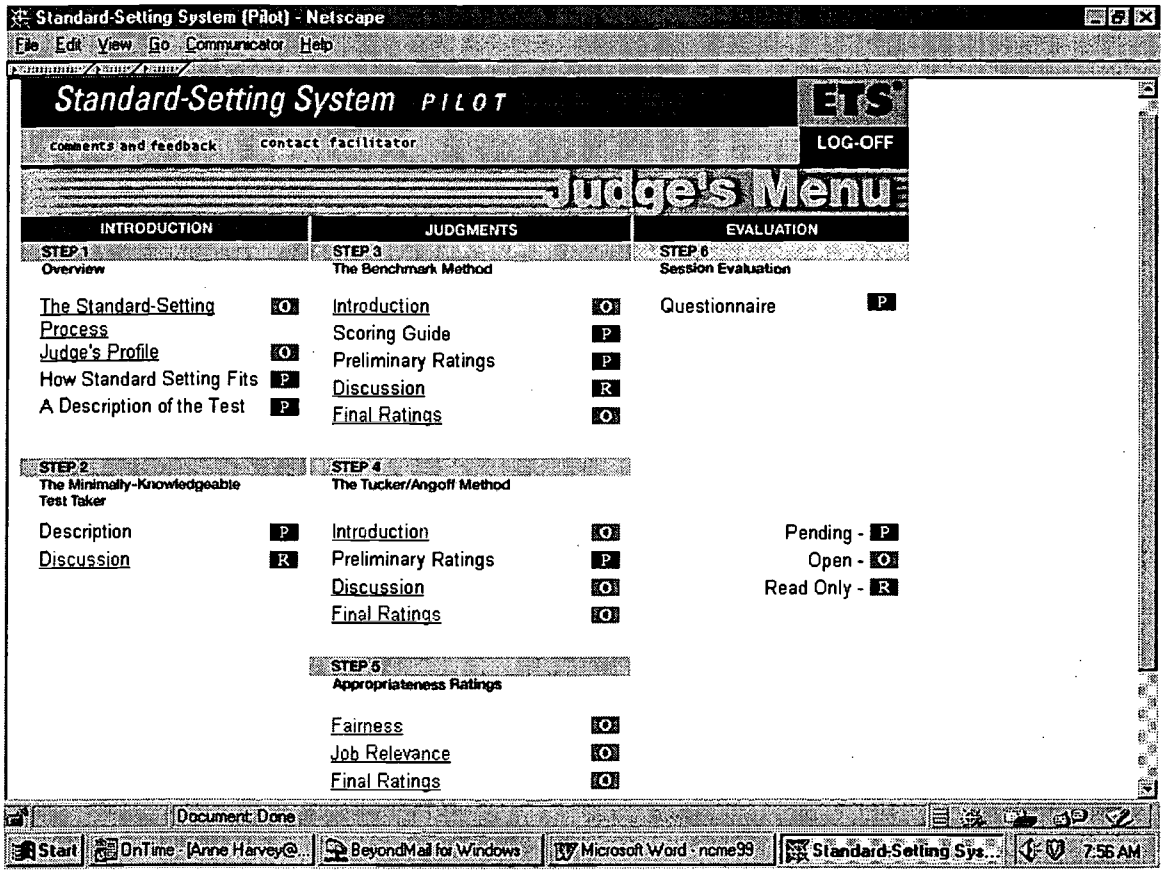
- Register and retire judges, allowing and disallowing access to the system
- View a particular judge's answers to the demographics and final questionnaires, time spent on each training page and item, judgements for each item, and comments made
- Allow access or deny access to any part of the study
- Close the study, disallowing any further changes to the data
- Participate in the discussions
- Compose messages to the judges
- Record comments for documentation
- Create data files
- Create summaries of the judgments for display to the judges

The judges have the following capabilities:

- Proceed through the training materials in the order specified
- Participate in the discussions
- Send e-mail to the facilitator
- Record comments
- Record and revise judgments

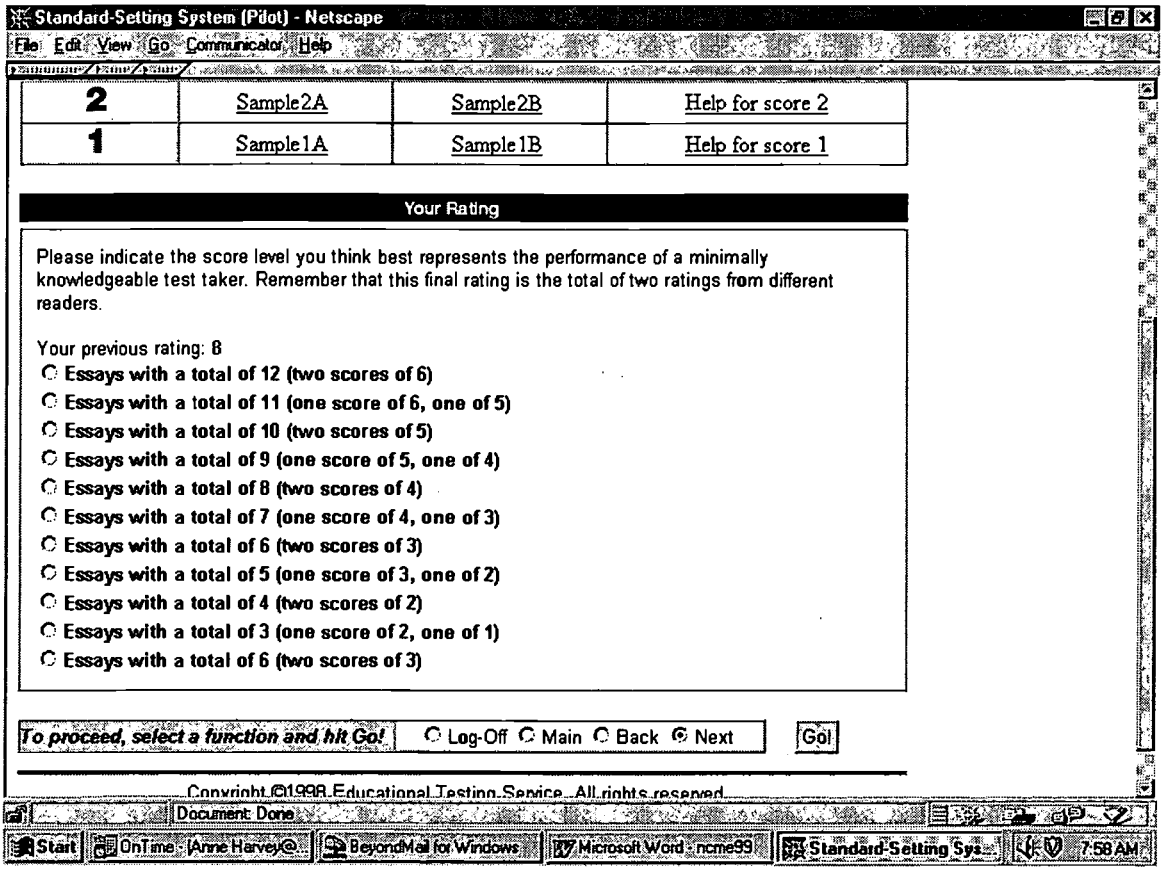
The prototype system begins with a welcome screen, which the facilitator can update with a new message as needed. Following the welcome screen the judge must agree to a non-disclosure statement before proceeding to the main menu (see Figure 1). Beyond this point, there is some flexibility in the workflow of a particular standard setting study. In the web-based portion of this study, the judge then proceeded through an introduction to standard setting and the test being studied.

Figure 1
Judges' Main Menu



The next step in the study was an introduction to, and discussion of, the minimally knowledgeable test taker. After sufficient discussion of the minimally knowledgeable test taker, the facilitator allowed the judges to proceed to training materials and preliminary judgments (see Figure 2) for the essay question. The benchmark method was used to determine a recommended passing score for the essay question.

Figure 2
Benchmark Rating Form

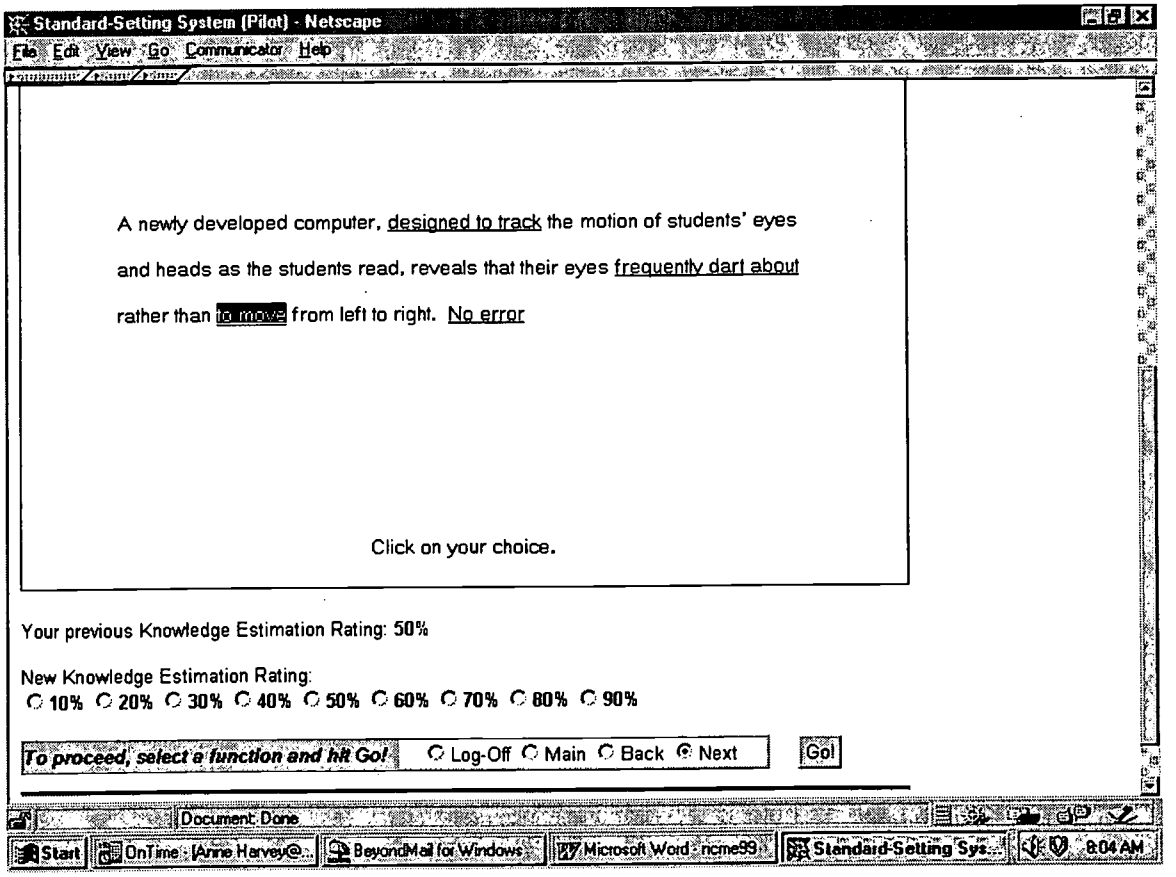


Preliminary judgments for the essay question were followed by a discussion. The discussion asked the judges to explain their ratings. Judges were allowed to make as many responses as they chose, and to respond to other judges' comments.

Once again, when the facilitator felt there had been sufficient discussion, the judges were allowed to proceed to the final ratings for the benchmark method. They were then able to proceed to the Angoff method training materials and preliminary Angoff ratings (see Figure 3). The preliminary ratings were completed for nine of the test items, three each from reading, writing, and

mathematics. Judges were shown each item and asked to answer it. When they proceeded to the next screen, the judges were shown both the item and the correct answer. They then were asked to make their judgment before proceeding to the next item.

Figure 3
Angoff Rating Form



Following the preliminary ratings, judges again participated in a discussion where they were asked to explain their ratings. When the facilitator felt there had been sufficient discussion, the judges were allowed to proceed to the final Angoff ratings. They were also allowed to proceed to training materials and ratings of fairness and job relevance of the items and test specifications.

Judges were then asked to complete the final questionnaire.

Methodology

Recommended passing scores were determined for a national teacher licensure test using both a monitored standard-setting study and a web-based study.

Participants

Participants were recruited from a list provided by the Department of Education in a southern state. Participants for the operational, monitored study were recruited from a list of experienced teachers with less than seven years experience by staff at the Department of Education. After recruitment was completed for the monitored study, participants for the Internet study were recruited from the same list. No teacher participated in both studies, and an attempt was made to match participants on the characteristics available. Those characteristics were the county in which they teach, teaching license level, and sex. A few teachers for each study were recruited based on recommendations from non-participating teachers. These teachers had more than seven years experience. Twenty-nine teachers were recruited for the operational study and twenty-one teachers were recruited for the Internet study.

Judges in the monitored study were allowed a substitute teacher for participating in the study, which took place during school hours. They were also reimbursed for their travel expenses, and in some cases for a hotel stay. Judges in the Internet study were paid a flat fee of $125 for their participation.

Materials

The test. The Praxis Computer-Based Academic Skills (CBAS) Assessment was used for this study. The CBAS Assessment tests the skills of students entering or exiting a teacher preparation program, including reading, writing, and mathematics. One constructed response item, a writing sample, is included in the assessment. All other items are multiple-choice items[2]. Because the assessment is computer adaptive, a representative set of items was selected to create a reference form for standard setting. The reference form contains 40 items each for reading, writing, and mathematics.

Biographical questionnaire. Participants in both studies filled out a biographical questionnaire. The questionnaire asked for the judges' teaching experience, racial/ethnic background, district type (urban, suburban, rural), highest degree obtained, and instructional level.

Final questionnaire. Participants in both studies filled out a final evaluation questionnaire at the end of the study. This questionnaire asked the judges to rate, on a scale from 1 (strongly disagree) to 5 (strongly agree), several aspects of the study. The statements the judges were asked to rate can be grouped into five categories: training, group process, general process, working environment, and navigation. There were small differences between the two questionnaires, such as changing "main menu" for the Internet study to "agenda" for the

[2] The multiple-choice items on this test are often presented without the traditional a, b, c, d, format. For example, test takers may be asked to choose a sentence with a particular characteristic by highlighting the sentence. The number of possible responses is sometimes large, but is limited to the choices presented. The items are, therefore, technically multiple choice rather than constructed response.

monitored study. Both groups also answered the question "Please rate your overall experience for the standard setting study (1=poor... ... 5=very good)."

Procedure

See Table 1 for a summary of the study events. This table lays out each step in the two studies sequentially. Some effort was made to provide a similar experience for the Internet and the monitored groups. However, several differences occurred. The differences can be placed in three categories: differences inherent to the Internet, differences due to choices available with the Internet, and differences due to working with a prototype system.

Table 1
Summary of the Study Events

	Panel-Based Study	Internet Study
Recruitment of judges	Recruited from a list of teachers in a southern state. The teachers taught at the elementary, middle, and high school level, and generally had less than seven years of teaching experience	Recruited from the same list of teachers as the panel-based study. An attempt was made to match teachers by teaching level, sex, and county in which they teach.
	Held in a centrally located hotel; ETS staff acted as facilitator	50% in their homes and 50% at their school; ETS staff acted as facilitator
Introduction	Registration	Assigned a user ID and password
	Welcome and introduction from ETS and state department of education	Welcome screen from ETS facilitator
	Background on test, purpose of the study, role of panel, facilitator, and state; review of test specifications	Background on test, purpose of the study, role of panel, facilitator, and state, review of test specifications
Minimally-knowledgeable Test Taker		Definition of minimally-knowledgeable test taker
		Discussion of minimally-knowledgeable test taker

Benchmark Method	Review scoring of Writing essay	Review scoring of Writing essay and definition of minimally knowledgeable teacher education candidate
	Define minimally knowledgeable test taker	
	Review benchmark responses and provide initial ratings	Review benchmark responses and provide initial ratings
	Summarize initial benchmark ratings and discuss in light of the minimally knowledgeable test taker	Summarize initial benchmark ratings and discuss in light of the minimally knowledgeable test taker
	Complete an additional round of ratings and discuss	
	Complete training evaluation form for benchmark method	Complete training evaluation form for benchmark method
	Complete final benchmark ratings	Complete final benchmark ratings
Pass/Fail Method	Complete initial pass/fail ratings for Writing essays	
	Summarize initial pass/fail ratings and discuss in light of the minimally knowledgeable test taker	
	Complete final pass/fail ratings	
	Train judges to rate test specifications for job relevance	
	Train judges to rate items for fairness	
	Review definition of minimally knowledgeable test taker	Review definition of minimally knowledgeable test taker
	Practice making Angoff judgments for multiple-choice items of different difficulty	Practice making Angoff judgments for multiple-choice items of different difficulty

	Make preliminary judgments for first five items of Writing test	Make preliminary judgments for nine selected items, three from each test
	Summarize initial multiple choice item ratings and discuss in light of the minimally knowledgeable test taker	Summarize initial multiple choice item ratings and discuss in light of the minimally knowledgeable test taker
	Complete training evaluation form for Angoff method	Complete training evaluation form for Angoff method
	Begin final Angoff ratings for Writing Test	Complete final Angoff ratings for all three tests
Job Relevance, Fairness, and Final Ratings	Complete training in making job relevance ratings for items	
	Complete job relevance ratings for first five items of Writing form	
	Summarize and discuss ratings of first five items	
	Complete Angoff, job relevance, and fairness ratings for Writing, Mathematics, and Reading	
		Complete training in making judgments of fairness
		Complete training in making judgments of job relevance
		Complete ratings of fairness and job relevance (both items and test specifications)
	Complete questionnaire	Complete questionnaire

Differences inherent to the Internet. While the facilitator for the monitored study was trained to particular standards for the testing program, some individual differences in training style will occur. Certain topics may be emphasized at the expense of others. The Internet study, on the other hand, was scripted with the input of several trained facilitators. The content of the training, therefore, was similar, but not necessarily identical.

One important difference is the ability to view items on a computer screen, rather than the written page. In this case, the test being analyzed is a computer-delivered test. The monitored sample viewed the items by looking at copies of screen prints. While the Internet sample viewed the items on a computer screen, the items were not exactly the same as they would be during the test. They were, in fact, pictures of the items as they would appear on the screen. The items could not be answered in the same way they would be when administered, and the size of the item could be slightly different.

The essays used as sample papers for the Internet sample when completing the benchmark method were also delivered as pictures on the computer, versus paper copies for the monitored sample. To the extent that there are paper-and-pencil test versus computer-based test differences for the test takers, there may be similar effects for the standard setting samples.

A final, obvious difference is that discussions took place face-to-face in the monitored study, but were in a written format for the Internet study.

Differences due to choices available with the Internet. For the Internet study, half of the participants chose to work at home and half chose to work at their school. The monitored study took place in a central location with a trained session facilitator.

The Internet study took approximately three weeks, with most participants spending about one hour per day, for fifteen of the twenty-one days. The monitored study took place over two days, approximately eight hours the first day and four hours the second day. There is nothing about the Internet that would prohibit the study from taking place over several hours of two days. However, most of the teachers participating in the study had access to their computers for only a few hours each day, either because of their work schedule or because the computer was shared with other teachers. There is also some question whether it would be prudent to require judges to spend eight hours in front of a computer screen in a given day. The fatigue factor could be considerable.

Some differences occurred in the order of the training. For example, the monitored study presented the Angoff method training, the item fairness training and the job relevance training before requiring judges to finish their ratings. This allowed judges who worked more quickly to leave when they were finished. The Internet study required judges to work sequentially, with the Angoff ratings being completed directly following the training.

Differences due to working with a prototype. The prototype web-based standard setting system used for this study was intended to determine the feasibility of using this method. Several options that would be available in a fully implemented system were not completed for the prototype:

- The monitored study used a pass/fail method in addition to the benchmark method for the writing essay; this method was not implemented in the Internet study.
- There were more frequent discussions for the monitored study than there were for the Internet study. The monitored study discussed the results after two rounds of preliminary rating, rather than the one round used for the Internet study. Also, preliminary job relevance and fairness ratings were discussed before completing final ratings for the monitored study, but not for the Internet study. Because of the importance of discussion in determining final ratings, and their peripheral nature to standard setting as usually implemented, the results of the fairness and job relevance ratings will not be discussed here.
- Software controls requiring complete data before proceeding were not implemented in the prototype. This led to some missing data, as indicated in the tables of results.

Results

Samples. Twenty-one teachers were recruited for the Internet study, 29 for the monitored study. The difference in the number recruited is illustrative of the recruiting effort, rather than the willingness of the teachers to participate. Approximately 75% of the teachers contacted agreed to participate, which is similar to the experience of the state's Department of Education in recruiting teachers for the monitored study, and of previous efforts to recruit for the testing program's studies. It is important to note, however, that an initial estimate of 7 hours to complete the Internet study proved quite unrealistic, with most participants taking around 15 hours to complete the study. This is reflected in the different completion rates for the two studies.

Of the 21 teachers recruited for the Internet study, 3 never began the study. This is similar to the 3 teachers who did not show up for the monitored study. However, of the 26 teachers who began the monitored study, all completed the study. Of the 19 teachers who began the Internet study, only 14 finished the study. One teacher indicated frustration with the slowness of the computer system she was working on. The other teachers indicated that the time for the study had exceeded their expectations and they could no longer participate.

Biographical information for the two study samples is presented in Table 2.

Table 2
Background and Experience of the Judges

	Internet Sample	Monitored Sample
Number Recruited	21	29
Number Not Arriving at (Beginning) Study	3 (14% of 21)	3 (10% of 29)
Number Not Completing Study	4 (19% of 21)	0 (0% of 29)
Total Number Completing Study	14	26
Sex:		
Men	3 (21%)	12 (46%)
Women	11 (79%)	14 (54%)
Ethnic Group		
White	13 (100%)	26 (100%)
Missing	1	0
Instructional Level		
Elementary	5 (42%)	7 (27%)
Middle	3 (25%)	3 (11%)
High School	4 (33%)	7 (27%)
Other	0 (0%)	9 (35%)[3]
Missing	2	0
District Type		
Urban	2 (15%)	4 (15%)
Suburban	3 (23%)	5 (19%)
Rural	8 (62%)	17 (65%)
Missing	1	0
Years of Teaching Experience		
1 to 5	6 (46%)	11 (42%)
6 to 10	6 (46%)	11 (42%)
11 to 15	0 (0%)	1 (4%)
16 to 20	1 (8%)	0 (0%
More than 20	0 (0%)	3 (12%)
Missing	1	0
Highest Degree		
B.A. or B.S.	9 (69%)	18 (69%)
M.A. or M.S.	4 (31%)	8 (31%)
Missing	1	0

[3] Reading Specialist, 2 Special Education, Art K-12, School Nurse, 2 Music K-12, Physical Education K-12, Counseling K-8

These results indicate that the two samples were similar in ethnic background, district type, teaching experience, and highest degree obtained. The Internet study panelists were more likely to be women and to be classroom teachers.

An alpha level of $p < .05$ was considered significant for all analyses of the two samples.

Final Questionnaire. The questionnaire was analyzed by averaging the responses for statements within each of the five categories, training, general process, group process, working environment, and navigation. These five scores were analyzed together with the overall experience question by completing a multivariate analysis of variance. The results are presented in Table 3.

The overall results for the MANOVA were statistically significant. Univariate follow-up tests indicated the results for the general process, working environment, and navigation statements were not significant. The general process statements referred to comfort in filling out the forms, understanding the purpose of each exercise, and confidence that the standard-setting process would produce a fair score. The working environment statements referred to the working conditions, any distractions, ease in getting to the work location, and compensation for effort. The navigation statements referred to navigation through the content, usefulness of the main menu or agenda, and whether the forms were easy to use.

Significant results were obtained for the training and group process scores. These two scores were further analyzed by completing a MANOVA for the statements within each of the categories. For the statement that asked whether the training was clear and complete, judges gave separate responses for the different parts of study. The responses to these statements were averaged for this analysis. The results from the training statements are presented in Table 4.

Table 3
Questionnaire Data

	Internet Sample (N = 13)	Monitored Sample (N = 26)
Training* (10 questions)		
Average (S.D.)	4.1 (.6)	4.6 (.4)
Range	3.1 to 5.0	3.8 to 5.0
Group Process* (4 questions)		
Average (S.D.)	3.5 (.5)	4.4 (.6)
Range	2.5 to 4.3	3.3 to 5.0
General Process (3 questions)		
Average (S.D.)	4.2 (.6)	4.4 (.6)
Range	3.0 to 5.0	3.0 to 5.0
Working Environment (4 questions)		
Average (S.D.)	4.2 (.8)	4.3 (.5)
Range	3.0 to 5.0	3.5 to 5.0
Navigation (6 questions)		
Average (S.D.)	4.2 (.5)	4.4 (.5)
Range	3.7 to 5.0	3.7 to 5.0
Overall Experience (1 question)		
Average (S.D.)	4.3 (.6)	4.6 (.6)
Range	3.0 to 5.0	3.0 to 5.0

*Statistically significant ($p < .05$)
Wilks' Lambda = .46 ($F = 6.19$, $p < .0002$)

Table 4
Questionnaire Data: Training

	Internet Sample (N = 13)	Monitored Sample (N = 26)
Training was clear and complete (6 questions) Average (S.D.) Range	4.2 (.4) 3.7 to 4.8	4.4 (.5) 3.0 to 5.0
I was comfortable asking questions* Average (S.D.) Range	4.0 (1.0) 2 to 5	4.7 (.5) 4 to 5
Questions were answered promptly* Average (S.D.) Range	3.8 (1.1) 2 to 5	4.7 (.5) 4 to 5
This was a good learning experience Average (S.D.) Range	4.6 (.5) 4 to 5	4.6 (.6) 3 to 5
Directions for the discussion were clear* Average (S.D.) Range	3.8 (1.0) 2 to 5	4.6 (.6) 3 to 5

*Statistically significant ($p < .05$)
Wilks' Lambda = .56 ($F = 5.26$, $p < .0012$)

The univariate follow-up analyses indicate significantly lower responses for the Internet sample for the statements about comfort asking questions, questions being answered promptly, and directions for participating in the discussion. There were no significant differences for statements regarding the clarity of the training materials or whether the study was a good learning experience.

The results for the group process statements are presented in Table 5.

Table 5
Questionnaire Data: Group Process

	Internet Sample (N = 13)	Monitored Sample (N = 26)
Discussions were helpful in rating items*		
Average (S.D.)	3.8 (.6)	4.4 (.6)
Range	2 to 4	3 to 5
Good opportunity to know colleagues and share ideas*		
Average (S.D.)	3.1 (1.3)	4.3 (.8)
Range	1 to 5	2 to 5
A good proportion of judges participated in the discussions*		
Average (S.D.)	3.5 (.7)	4.3 (.9)
Range	2 to 4	2 to 5
I was comfortable sharing my ideas with other judges*		
Average (S.D.)	3.8 (.4)	4.6 (.5)
Range	3 to 4	4 to 5

*Statistically significant ($p < .05$)
Wilks' Lambda = .56 ($F = 6.74$, $p < .0004$)

In this case, responses to all of the statements were significantly different, with the Internet sample scoring lower each time. Two of the Internet sample judges commented that it was difficult to answer the question about participation of other judges, as they did not have a clear idea of how many judges were participating in the study.

Benchmark method. The benchmark method results were analyzed using a repeated measures analysis of variance with sample (Internet or monitored) as the independent variable, time (preliminary and final judgments) as the repeated measure and rating as the dependent variable. The preliminary judgments were significantly different from each other, as were the final ratings (see Tables 6 and 7). The interaction, however, was not significant, indicating that the different modes of discussion did not have a differential effect on the results.

Table 6
Benchmark Judgments

	Internet Sample (N = 12)	Monitored Sample (N = 26)	Praxis Historic Range (8 studies)
First Judgment Average (S.D.) Range	7.0 (1.59) 4 to 10	8.0 (1.26) 4 to 10	
Final Judgment Average (S.D.) Range	6.9 (.90) 6 to 9	7.9 (.93) 6 to 9	7.3 6.8 to 8.0
Difference Between Initial and Final Judgments	.1	.1	

Table 7
Repeated Measures Analysis of Variance for the Benchmark Method

Source	DF	Sum of Squares	Mean Square	F	Pr>F
Sample	1	16.5265	16.5265	8.00	.0076
Error (Sample)	36	74.3814	2.0662		
Time	1	.1054	.1054	.16	.6955
Time x Sample	1	.0002	.0002	.00	.9875
Error (Time)	36	24.3814	.6773		

As shown in Table 6, eight previous standard setting studies for the same test were summarized to determine whether the final results for this study were within the historic range for the testing program. Both samples were within the historic range.

If we compare the results of the two samples by taking the ratio of the highest standard to the lowest standard (Jaeger, 1989), we find a ratio of 1.14 for the essay.

Angoff method results. The Angoff method results were analyzed using a MANOVA, with sample as the independent variable and average reading, writing, and mathematics ratings as the dependent variables. The preliminary results

were not analyzed, as different subsets of items were used for the two groups for the preliminary ratings. The results of the MANOVA were not significant. The results are presented in Table 8.

Table 8
Angoff Judgments

	Internet Sample (N = 14)	Panel Sample (N = 26)	Difference	Praxis Historical Ranges (5 studies)
Reading Average (S.D.) Range	25.4 (3.1) 19.5 to 30.6	25.6 (3.8) 13.7 to 31.3	-0.2	25.1 (2.0) 22.0 to 27.4
Writing Average (S.D.) Range	26.0 (2.8) 21.9 to 29.9	24.2 (3.0) 15.0 to 28.7	1.8	24.3 (.8) 22.9 to 25.0
Mathematics Average (S.D.) Range	24.5 (2.4) 20.3 to 27.8	23.9 (4.1) 14.5 to 32.8	0.6	23.2 (.9) 21.9 to 24.2

Wilks' Lambda = .82 ($F = 2.64$, $p < .0645$)

As with the benchmark method, the results for five previous standard setting studies for the same tests were summarized to determine whether the final results for this study were within the historic range for the testing program. The recommended passing scores for the reading test for both samples were well within the historic range. The results for the writing and mathematics tests for the monitored sample were within the historic range for the program, but results for the Internet sample were slightly higher than the range.

If we compare the results of the two samples by taking the ratio of the highest standard to the lowest standard, we find a ratio of 1.01 for reading, 1.07 for writing, and 1.03 for mathematics.

Discussion

The results of this study are generally favorable toward using a web-based standard setting system. While the results of the benchmark method are significantly different for the Internet and the monitored samples, the greatest concern, that the discussion will not have the intended effect when completed over the Internet, was not substantiated. As is typical for iterative studies of this nature, the recommended passing scores stayed roughly the same (lower by .1 on a scale of 2 to 12 for both samples), but the standard deviations of the judges'

ratings were substantially reduced. In fact, the standard deviation for the Internet group was reduced to a greater degree than that of the monitored group.

The fact that the initial results for the benchmark method are significantly different is of some concern. However, despite some effort to obtain similar samples and to use similar training materials, there were many differences in the final design of the two studies. For example, two of the 12 (17%) Internet study judges had a degree in English while 3 of the 26 (12%) monitored study judges had a degree in English. There may also have been a gender effect, with a considerably higher percentage of the Internet sample than the monitored sample being female.

Another difference is that the Internet sample had a more formal discussion of the minimally knowledgeable test taker than did the monitored sample prior to the preliminary ratings. It might be expected, however, that this difference would wash out after the later discussions, which was not the case. The difference in the recommended passing scores was the same after the discussions as it was before the discussions.

Although it was considered more important to take advantage of the relative benefits of the Internet and monitored sample methods in comparing the two, future studies might reduce the unnecessary differences, such as the non-random nature of the samples. It will be important to discover whether a lower recommended passing score initially is due to non-random differences in the sample, the training materials, the method of reviewing sample papers, or some other factor. In the mean time, the results for both samples would be considered appropriate given the historic ranges for the program.

Although the Angoff results for the Internet sample were slightly higher than the historical ranges for two of the three tests, they were not significantly different from the results of the monitored sample. In his 1989 chapter on certification of student competence, Jaeger compared different methods by taking the ratio of the highest standard to the lowest standard. In comparing different methods, Jaeger found ratios from 1.00 to 52.00, with a median value of 1.46 and an average value of 5.30. If we go so far as to assume the web-based Angoff and the monitored Angoff are different methods, ratios of 1.01 for reading, 1.07 for writing, and 1.03 for math suggest little cause for concern.

The suspicion that teachers would not feel as positive about the group process was born out by the significantly lower scores given by the Internet sample to statements about the discussions' helpfulness in rating items and whether or not the study was a good opportunity to get to know colleagues. It was surprising, however, to find that teachers were less comfortable sharing their ideas over the Internet than they were in a face-to-face meeting. While this question was intended to elicit responses about the "bully" factor in discussions, it may in fact have more to do with the judge's comfort in the mode of responding. This is evidenced by the significantly lower scores for the Internet sample when

responding to the statement "The directions for participating in the discussion were clear."

In future versions of the web-based system it might be helpful to expand on the directions for the discussion, including providing samples of some previous responses. It may also be helpful to provide the judges with more information about each other, including the number of judges participating in the study and a little of their background.

An experienced facilitator[4], in reviewing the Internet discussion, noted that the responses seemed qualitatively different from the discussions that typically take place in monitored sessions. The speculation is that writing a response, rather than saying a response, may result in a different response. It would be interesting to compare a transcript of a monitored discussion with that same discussion over the Internet. Some interesting comparisons might be the length of the response, the number of contradictory responses, and the number of repeated phrases.

It was also not surprising that judges in the Internet sample did not feel their questions were answered as promptly. Although several methods were provided for contacting the facilitator (pager, e-mail, a 'contact facilitator' button, and a 'comments' button), nothing could be as immediate as a face-to-face session. In fact, the judges may have been somewhat confused about which method of contacting the facilitator to use in a given situation. The instructions might be simplified in future releases of the system.

What is most encouraging is that judges in both samples responded similarly to statements about the general process, including a question about the fairness of the results, and about the experience as a whole. It remains to be seen whether judges can be as easily recruited when told the study will take 15 hours as they were when told 7 hours. However, the equipment for the study was apparently available to the teachers, and they were, for the most part, interested in participating via this method.

In sum, the web-based standard setting system appears to be a method worth developing. A future version of the system should improve on the communication between the judges, and between the judges and facilitator. However, the teachers in this study felt similarly about the overall experience, regardless of whether it was an Internet or a face-to-face study. The results of this study suggest that recommended passing scores from an Internet study will be similar to those from a monitored study.

[4] Jane Faggen, personal communication, November 1998

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thomdike (Ed.), *Educational Measurement*. Washington, DC: American Council on Education.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research*, *56*(1), 137-172.

Cizek, G. J. & Fitzgerald, S. M. (1996, April). *A Comparison of Group and Independent Standard Setting*. Paper presented at the Annual Meeting of the American Educational Research Association: New York, NY.

Crocker, L. & Zieky, M. (1995). *Joint Conference on Standard Setting for Large-Scale Assessments (Washington, D.C., October 5-7, 1994): Executive Summary, Volume I.* National Center for Education Statistics: Washington, D.C.

Faggen, J. (1994, November). *Setting Standards for Constructed-Response Tests: An Overview.* Research Memorandum RM-94-19. Princeton, NJ: Educational Testing Service.

Fitzpatrick, A. R. (1989). Social influences in standard setting: The effects of social interaction on group judgements. *Review of Educational Research*, *59*(3), 315-328.

Hambleton, R. K. & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education, 8*(1), 41-55.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 485-514). New York: American Council on Education and Macmillan.

Plake, B. S. & Impara, J. C. (1996). *Intrajudge Consistency Using the Angoff Standard-Setting Method.* Paper presented at the Annual Meeting of the National Council on Measurement in Education: New York, NY.

Authors' Notes

Anne L. Harvey is a Measurement Statistician II in the Analysis Division at Educational Testing Service. Walter D. Way is Executive Director for the Analysis Division at Educational Testing Service.

The authors would like to thank several individuals, without whom this study would not have been possible. Ken Berger, Ted Blew, and Dick Bohlander were instrumental in the design of the prototype web-based system. Ann Angell, Dick Bohlander, Jane Faggen, Kathy Martin, Robert Smith, and Michael Zieky made several significant contributions while reviewing the training materials for the web-based system. Dick Bohlander, Richard Carvalho, and Candace Mero did an excellent job of recruiting for the Internet portion of the study. Dan Jacquemin helped with the transfer of the items from paper to the web. Nancy Feryok assisted with the data analysis. Reviews of earlier drafts of this paper by Jane Faggen and J. T. Stewart resulted in some very helpful changes. J. T. Stewart acted as the facilitator for the monitored study. Candace Mero was extremely helpful in checking the accuracy of the results in this report.

Requests for additional copies of this report may be directed to Anne Harvey, Mailstop 15-L, Educational Testing Service, Princeton, NJ 08541-0001

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: A COMPARISON OF WEB_BASED STANDARD SETTING AND MONITORED STANDARD SETTING	
Author(s): ANNE L. HARVEY	
Corporate Source: EDUCATIONAL TESTING SERVICE	Publication Date: APRIL 1999

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Anne L. Harvey	Printed Name/Position/Title: ANNE L. HARVEY, DIR. OF PRODUCT DEV.
Organization/Address: THE COLLEGE BOARD, 45 COLUMBUS AVENUE, NY, NY 10023	Telephone: 212-713-8070 FAX: 212-649-8427
	E-Mail Address: aharvey@collegeboard.org Date: 3/13/02

(over)

ERIC®

Clearinghouse on Assessment and Evaluation

University of Maryland
1129 Shriver Laboratory
College Park, MD 20742-5701

Tel: (800) 464-3742
(301) 405-7449
FAX: (301) 405-8134
ericae@ericae.net
http://ericae.net

May 8, 2000

Dear AERA Presenter,

Hopefully, the convention was a productive and rewarding event. As stated in the AERA program, presenters have a responsibility to make their papers readily available. If you haven't done so already, please submit copies of your papers for consideration for inclusion in the ERIC database. We are interested in papers from this year's AERA conference and last year's conference. If you have submitted your paper, you can track its progress at http://ericae.net.

Abstracts of papers accepted by ERIC appear in *Resources in Education (RIE)* and are announced to over 5,000 organizations. The inclusion of your work makes it readily available to other researchers, provides a permanent archive, and enhances the quality of *RIE*. Abstracts of your contribution will be accessible through the printed and electronic versions of *RIE*. The paper will be available through the microfiche collections that are housed at libraries around the world and through the ERIC Document Reproduction Service.

We are gathering all the papers from the **2000 and 1999 AERA Conference**. We will route your paper to the appropriate clearinghouse. You will be notified if your paper meets ERIC's criteria for inclusion in *RIE*: contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality.

Please sign the Reproduction Release Form enclosed with this letter and send **two** copies of your paper. The Release Form gives ERIC permission to make and distribute copies of your paper. It does not preclude you from publishing your work. You can mail your paper to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to: AERA 2000/ERIC Acquisitions
University of Maryland
1129 Shriver Laboratory
College Park, MD 20742

Sincerely,

Lawrence M. Rudner

Lawrence M. Rudner, Ph.D.
Director, ERIC/AE